What is claimed is:

1. A system for finding compounds in a text corpus, comprising:

a vocabulary comprising tokens extracted from a text corpus; and

a compound finder iteratively identifying compounds having a plurality of lengths within the text corpus, each compound comprising a plurality of tokens, comprising:

an *n*-gram counter evaluating a frequency of occurrence for one or more *n*-grams in the text corpus, each *n*-gram comprising tokens selected from the vocabulary; and

a likelihood evaluator determining a likelihood of collocation for one or more of the *n*-grams having a same length, adding the *n*-grams having a highest likelihood as compounds to the vocabulary and rebuilding the vocabulary based on the added compounds.

2. A system according to Claim 1, further comprising:

an iterator selecting *n*-grams having a same length that is less than the same length as *n*-grams previously selected.

3. A system according to Claim 1, wherein only some of the *n*-grams having a highest likelihood are added as compounds to the vocabulary.

4. A system according to Claim 1, wherein the likelihood of collocation as a likelihood ratio λ is computed in accordance with the formula:

$$\lambda = \frac{L(H_i)}{L(H_c)}$$

where $L(H_i)$ is a likelihood of observing H_i under an independence hypothesis, $L(H_c)$ is a likelihood of observing H_c under a collocation hypothesis, and H is a pair of tokens.

5. A system according to Claim 4, wherein the $L(H_c)$ is determined, comprising dividing the *n*-gram into *n*-1 pairings of segments, calculating a

likelihood of collocation for each pairing of segments, and selecting the maximum likelihood of collocation of the pairings as $L(H_c)$.

6. A method for finding compounds in a text corpus, comprising:

building a vocabulary comprising tokens extracted from a text corpus; and

iteratively identifying compounds having a plurality of lengths within the text corpus, each compound comprising a plurality of tokens, comprising:

evaluating a frequency of occurrence for one or more *n*-grams in the text corpus, each *n*-gram comprising tokens selected from the vocabulary;

determining a likelihood of collocation for one or more of the *n*-grams having a same length; and

adding the *n*-grams having a highest likelihood as compounds to the vocabulary and rebuilding the vocabulary based on the added compounds.

7. A method according to Claim 6, further comprising:

selecting *n*-grams having a same length that is less than the same length as *n*-grams previously selected.

8. A method according to Claim 6, further comprising:

adding only some of the *n*-grams having a highest likelihood as compounds to the vocabulary.

9. A method according to Claim 6, further comprising:

computing the likelihood of collocation as a likelihood ratio λ in accordance with the formula:

$$\lambda = \frac{L(H_i)}{L(H_c)}$$

where $L(H_i)$ is a likelihood of observing H_i under an independence hypothesis, $L(H_c)$ is a likelihood of observing H_c under a collocation hypothesis, and H is a pair of tokens.

10. A method according to Claim 9, further comprising:

determining $L(H_c)$, comprising:

dividing the n-gram into n-1 pairings of segments;

calculating a likelihood of collocation for each pairing of segments; and

selecting the maximum likelihood of collocation of the pairings as $L(H_c)$.

11. A computer-readable storage medium holding code for performing the method according to Claim 6.

12. An apparatus for finding compounds in a text corpus, comprising:

means for building a vocabulary comprising tokens extracted from a text corpus; and

means for iteratively identifying compounds having a plurality of lengths within the text corpus, each compound comprising a plurality of tokens, comprising:

means for evaluating a frequency of occurrence for one or more n-grams in the text corpus, each n-gram comprising tokens selected from the vocabulary;

means for determining a likelihood of collocation for one or more of the n-grams having a same length; and

means for adding the n-grams having a highest likelihood as compounds to the vocabulary and means for rebuilding the vocabulary based on the added compounds.

13. A system for identifying compounds through iterative analysis of measure of association, comprising:

a stored limit on a number of tokens per compound; and

a compound finder iteratively evaluating compounds within a text corpus, comprising:

an n-gram counter determining a number of occurrences of one or more n-grams within the text corpus, each n-gram comprising up to a maximum number of tokens, which are each provided in a vocabulary for the text corpus;

a likelihood evaluator identifying at least one n-gram comprising a number of tokens equal to the limit based on the number of occurrences and determining a measure of association between the tokens in the identified n-gram and adding each identified n-gram with a sufficient measure of association to the vocabulary as a compound token, rebuilding the vocabulary based on the added compound tokens and adjusting the limit.

14. A system according to Claim 13, further comprising:

a stored upper limit on a number of identified n-grams; and

a limiter identifying a number of n-grams up to the upper limit based on the number of occurrences.

15. A system according to Claim 13, further comprising:

an iterator initially specifying the limit comprising a plurality of tokens per compound and subsequently decreasing the limit comprising a lesser plurality of tokens per compound.

16. A system according to Claim 13, wherein the measure of association between the tokens in the identified n-gram comprises a likelihood ratio λ.

17. A system according to Claim 16, wherein the likelihood ratio λ is calculated in accordance with the formula:

$$\lambda = \frac{L(H_i)}{L(H_c)}$$

where $L(H_i)$ is a likelihood of observing H_i under an independence hypothesis, $L(H_c)$ is a likelihood of observing H_c under a collocation hypothesis, and H is a pair of tokens.

18. A system according to Claim 17, wherein, for each pair of tokens, t_1, t_2, in the identified n-gram, the independence hypothesis comprises $P(t_2 \mid t_1) = P(t_2 \mid \overline{t_1})$ and the collocation hypothesis comprises $P(t_2 \mid t_1) > P(t_2 \mid \overline{t_1})$.

19. A system according to Claim 17, wherein the $L(H_i)$ is computed for each pair of tokens, t_1, t_2, in the identified n-gram in accordance with the formula:

$$\arg\max_{L(H_i)} \frac{L(t_1, t_2 \, form \, compound)}{L(n-gram \, does \, not \, form \, compound)}.$$

20. A system according to Claim 13, further comprising:

an initial vocabulary comprising a plurality of tokens extracted from the text corpus.

21. A system according to Claim 20, further comprising:

a parser parsing the tokens from the text corpus.

22. A system according to Claim 13, further comprising:

a filter determining the number of occurrences of one or more n-grams within the text corpus for only unique n-grams.

23. A system according to Claim 13, wherein each text corpus comprises a plurality of documents comprising one of a Web page, a news message and text.

24. A method for identifying compounds through iterative analysis of measure of association, comprising:

specifying a limit on a number of tokens per compound; and

iteratively evaluating compounds within a text corpus, comprising:

determining a number of occurrences of one or more n-grams within the text corpus, each n-gram comprising up to a maximum number of tokens, which are each provided in a vocabulary for the text corpus;

identifying at least one n-gram comprising a number of tokens equal to the limit based on the number of occurrences and determining a measure of association between the tokens in the identified n-gram; and

adding each identified n-gram with a sufficient measure of association to the vocabulary as a compound token, rebuilding the vocabulary based on the added compound tokens and adjusting the limit.

25. A method according to Claim 24, further comprising:

providing an upper limit on a number of identified n-grams; and

identifying a number of n-grams up to the upper limit based on the number of occurrences.

26. A method according to Claim 24, further comprising:

initially specifying the limit comprising a plurality of tokens per compound; and

subsequently decreasing the limit comprising a lesser plurality of tokens per compound.

27. A method according to Claim 24, wherein the measure of association between the tokens in the identified n-gram comprises a likelihood ratio λ.

28. A method according to Claim 27, further comprising:

calculating the likelihood ratio λ in accordance with the formula:

$$\lambda = \frac{L(H_i)}{L(H_c)}$$

where $L(H_i)$ is a likelihood of observing H_i under an independence hypothesis, $L(H_c)$ is a likelihood of observing H_c under a collocation hypothesis, and H is a pair of tokens.

29. A method according to Claim 28, wherein, for each pair of tokens, t_1, t_2, in the identified n-gram, the independence hypothesis comprises $P(t_2 \mid t_1) = P(t_2 \mid \overline{t_1})$ and the collocation hypothesis comprises $P(t_2 \mid t_1) > P(t_2 \mid \overline{t_1})$.

30. A method according to Claim 28, further comprising:

computing the $L(H_i)$ for each pair of tokens, t_1, t_2, in the identified n-gram in accordance with the formula:

$$\arg\max_{L(H_i)} \frac{L(t_1, t_2 \, form \; compound)}{L(n-gram \; does \; not \; form \; compound)}.$$

31. A method according to Claim 24, further comprising:

constructing an initial vocabulary comprising a plurality of tokens extracted from the text corpus.

32. A method according to Claim 31, further comprising:

parsing the tokens from the text corpus.

33. A method according to Claim 24, further comprising:

determining the number of occurrences of one or more n-grams within the text corpus for only unique n-grams.

34. A method according to Claim 24, wherein each text corpus comprises a plurality of documents comprising one of a Web page, a news message and text.

35. A computer-readable storage medium holding code for performing the method according to Claim 24.

36. An apparatus for identifying compounds through iterative analysis of measure of association, comprising:

means for specifying a limit on a number of tokens per compound; and

means for iteratively evaluating compounds within a text corpus, comprising:

means for determining a number of occurrences of one or more *n*-grams within the text corpus, each *n*-gram comprising up to a maximum number of tokens, which are each provided in a vocabulary for the text corpus;

means for identifying at least one *n*-gram comprising a number of tokens equal to the limit based on the number of occurrences and means for determining a measure of association between the tokens in the identified *n*-gram; and

means for adding each identified *n*-gram with a sufficient measure of association to the vocabulary as a compound token, means for rebuilding the vocabulary based on the added compound tokens and means for adjusting the limit.